Hourly surface nitrogen dioxide retrieval from GEMS tropospheric vertical column densities: benefit of using time-contiguous input features for machine learning models

J. Gödeke[2], A. Richter[1], K. Lange[1], P. Maaß[2]

[1]Institute of Environmental Physics
University of Bremen, Germany
Email: Andreas.Richter@iup.physik.uni-bremen.de

[2]Center for Industrial Mathematics,
University of Bremen, Bremen, Germany

Aim of the study

- Nitrogen oxides (NOx = NO_2 + NO) are important trace gases in the troposphere.
- They are a key player in tropospheric ozone formation.
- NO_2 adversely affects human health.

- Satellite observations provide tropospheric NO_2 columns, which are linked to NO_2 surface concentrations.
- The Korean geostationary GEMS instrument was the first to provide hourly NO_2 columns.
- Here, we use the IUP-UB GEMS tropospheric NO_2 product.

- ML techniques can be used to derive surface concentrations from satellite columns.
- Hourly observations provide additional information to improve ML predictions.

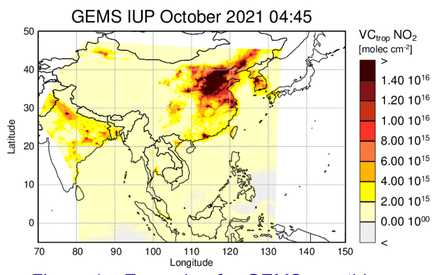

Figure 1: Example of a GEMS monthly mean tropospheric NO_2 field

Data used

Satellite Data
- GEMS IUP-UB NO_2 columns
- January 2021 - November 2022
- over South Korea

Meteorological Data
- ERA5, see table

Other Data
- latitude
- height
- soil type
- vegetation

Feature name	Source
Tropospheric vertical column density of NO_2	IUP-UB retrieval on GEMS data
Latitude at the center of GEMS pixel	GEMS data product
Surface height at the center of GEMS pixel	GEMS data product
10 m u component of wind	ERA5
100 m u component of wind	ERA5
Instantaneous 10 m wind gust	ERA5
2 m temperature	ERA5
Surface pressure	ERA5
Skin temperature	ERA5
UV–visible albedo for diffuse radiation	ERA5
Downward UV radiation at the surface	ERA5
UV–visible albedo for direct radiation	ERA5
Boundary layer height	ERA5
Total column water	ERA5
Evaporation	ERA5
Soil type	ERA5
High vegetation cover	ERA5

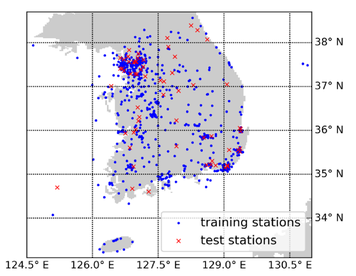

Figure 2: An exemplary split into 90% training stations and 10% test stations, considered during multiple 10-fold spatial cross-validations.

Validation Data
- hourly data from 637 surface in-situ stations
- no selection for type of station (urban, rural, background)
- 60 random splits into training (90%) and test (10%) data
- averaging of test results

Data handling and hyperparameters

Pairing:
Satellite and other feature data are matched to in-situ observations in time and space.

Selection:
Satellite data are selected for quality and cloud free scenes. This results in many gaps in time and space.

Time contiguous data sets:
Data sets are constructed, which consist of data having at least (k - 1) previous hours of measurements and contain N elements. Larger k implies lower N as not all data points have data for earlier hours.

Linear regression:
- not shown, see publication for results

Random forest:
- Python scikit-learn software package
- `max_features` = 2, 3, 3, 3, 4 for time contiguity k = 1, 2, 3, 4, 5
- `min_samples_leaf` and `max_samples` = 5 using 100% of the size of the training data
- `n_estimators` = 8000 trees (not needed but used to improve stability)
- All remaining hyperparameters are always set to the default values

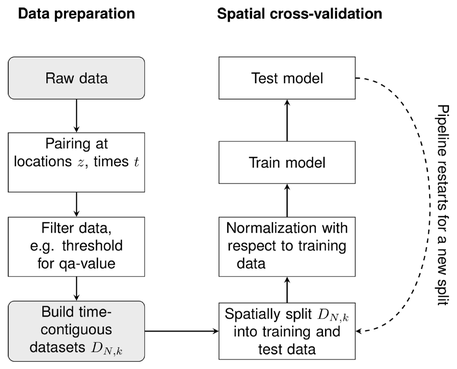

Figure 3: Flowchart for all data processing steps. The left column shows the construction of the time-contiguous datasets $D_{N,k}$. Evaluating the performance of models on $D_{N,k}$ is done via spatial cross-validation (right column).

Results

1) Does higher data contiguity improve the results?

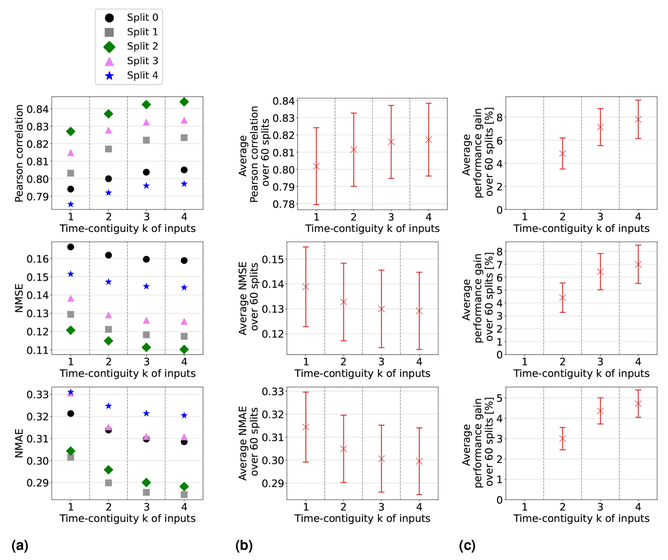

Figure 4: Random forests trained and tested on datasets $D_{4,k}$ for 60 different splits into training and test stations, with different time contiguity k of the input features. In panel (a), performances on test sets are shown for five exemplary station splits with respect to three performance measures. Panel (b) shows the average performance over all 60 splits, with error bars illustrating the standard deviation. Panel (c) shows the average performance gain relative to the case k = 1. Across each row, the same performance measure is considered.

2) Does higher data contiguity outweigh loss of data points?

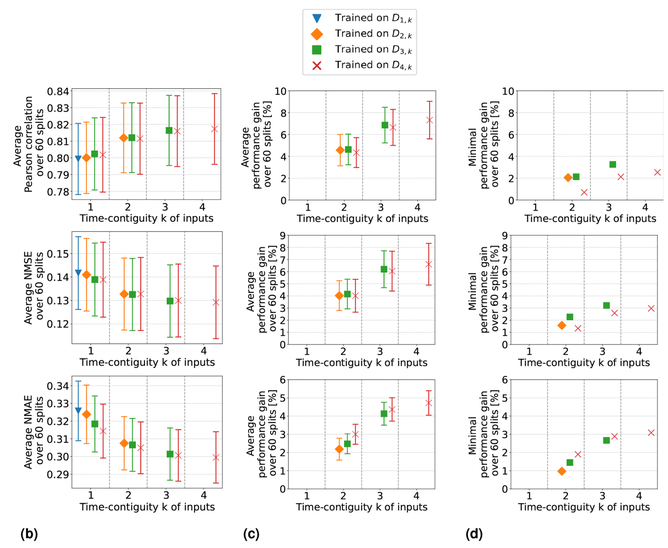

Figure 5: Random forests trained on $D_{M,k}$ for M < 4 with different time contiguities k. Performance on $D_{4,k}$ has been evaluated through six-times 10-fold spatial cross-validation. Panel (a) shows the average performance over all 60 station splits for three performance measures. Panel (b) shows the average performance gain relative to the best case of k = 1. Error bars illustrate the standard deviation. Panel (c) shows the minimal performance gain. Across each row the same performance measure is considered.

3) What is the impact of satellite data, latitude and surface height?

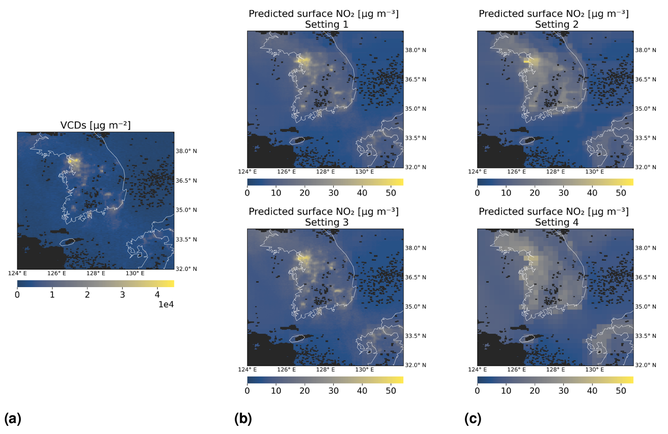

Figure 5: Predictions of surface NO_2 by random forests on 7 April 2021 from 01:00 to 02:00 UTC. Panel (a) shows tropospheric NO_2 VCDs from 01:45 to 02:15 UTC. Panel (b) shows predicted surface NO_2 in Settings 1 (all features) and 3 (latitude and height excluded). Panel (c) shows predictions in Settings 2 (no satellite data) and 4 (no satellite data, no latitude, no surface height). The black mask indicates missing data, e.g., due to clouds. All models have been trained with time contiguity k = 4 on $D_{4,4}$ for the same choice of training stations.

Conclusions

- Hourly tropospheric NO_2 columns from the GEMS satellite instrument were used to predict surface NO_2 concentrations with a random forest model.
- Including time contiguous measurements from earlier times in the training and prediction always improved predictions relative to surface in-situ measurements.
- Performance gains are also observed when considering the lower number of time contiguous data points available.
- In practical applications, for each point, the maximum available time contiguity should be applied up to k = 4. Larger k does not result in further improvements.
- Excluding satellite NO_2 observations from the feature list worsened the predictions but was still in an acceptable range. However, it is expected that the model without satellite data does not perform as well outside of South Korea.
- Application to other geostationary instruments, such as TEMPO and Sentinel-4, will be interesting.

Acknowledgements

- We thank the National Institute of Environmental Research (NIER) of South Korea for providing GEMS lv1 data and financial support (NIER-2022-04-02-037)
- in situ measurements of surface NO_2 were provided by the Korean Ministry of Environment
- This project has been funded by the Deutsches Zentrum für Luft- und Raumfahrt (grant no. 50 EE 2204)

Selected references

Gödeke, J., Richter, A., Lange, K., Maaß, P., Hong, H., Lee, H., and Park, J.: Hourly surface nitrogen dioxide retrieval from GEMS tropospheric vertical column densities: benefit of using time-contiguous input features for machine learning models, Atmos. Meas. Tech., 18, 3747–3779, https://doi.org/10.5194/amt-18-3747-2025, 2025.

Lange, K., Richter, A., Bösch, T., Zilker, B., Latsch, M., Behrens, L. K., Okafor, C. M., Bösch, H., Burrows, J. P., Merlaud, A., Pinardi, G., Fayt, C., Friedrich, M. M., Dimitropoulou, E., Van Roozendael, M., Ziegler, S., Ripperger-Lukosiunaite, S., Kuhn, L., Lauster, B., Wagner, T., Hong, H., Kim, D., Chang, L.-S., Bae, K., Song, C.-K., Park, J.-U., and Lee, H.: Validation of GEMS tropospheric NO2 columns and their diurnal variation with ground-based DOAS measurements, Atmos. Meas. Tech., 17, 6315–6344, https://doi.org/10.5194/amt-17-6315-2024, 2024.

University of Bremen

see also: www.doas-bremen.de